EUROPEAN PATENT OFFICE

Patent Abstracts of Japan

PUBLICATION NUMBER : 11282492
PUBLICATION DATE : 15-10-99

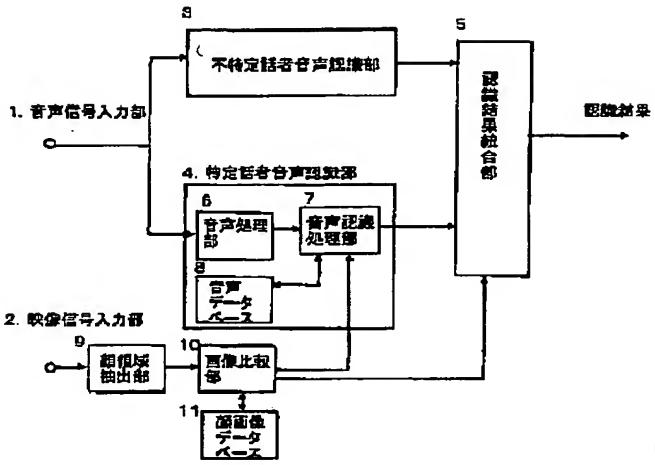
APPLICATION DATE : 26-03-98
APPLICATION NUMBER : 10079916

APPLICANT : MATSUSHITA ELECTRIC IND CO LTD;

INVENTOR : INOUE IKUO;

INT.CL. : G10L 3/00 G10L 3/00 G10L 3/00 G06T 1/00 G10L 5/06

TITLE : SPEECH RECOGNITION DEVICE, SPEAKER DETECTOR, AND IMAGE RECORDER



ABSTRACT : PROBLEM TO BE SOLVED: To achieve speech recognition of plural speakers with high reliability.

SOLUTION: This system is configured of a speech signal input part 1, a video signal input part 2, an unspecified speaker speech recognition part 3 for extracting a common feature from speeches of multi-speakers, making a standard pattern, and calculating a degree of similarity between the input speeches and a standard speech pattern, a specific speaker speech recognition part 4 for calculating a degree of similarity between the input speech and the speech of a pre-registered speaker, a face region extracting part 9 for extracting a face region from an input video, a face image database 11 for recording face image data of plural specific speakers and their identification numbers, an image comparison part 10 for outputting the degree of similarity with the image data inputted from the face region extracting part 9 and the face image database 11, and a recognition result integration part 5 for calculating an integrated degree of similarity from the outputs of the unspecified speaker speech recognition part 3, the specific speaker speech recognition part 4, and the image comparison part 10, and outputting the recognition result.

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－282492

(43)公開日　平成11年(1999)10月15日

(51)Int.Cl.[6]		識別記号	FI		
Ｇ１０Ｌ	3/00	５７１	Ｇ１０Ｌ	3/00	５７１Ｃ
		５１３			５１３Ｚ
		５３１			５３１Ｊ
					５３１Ｋ
Ｇ０６Ｔ	1/00			5/06	Ｄ

審査請求　未請求　請求項の数12　ＯＬ　（全 11 頁）　最終頁に続く

(21)出願番号　特願平10－79916

(22)出願日　平成10年(1998) 3月26日

(71)出願人　000005821
松下電器産業株式会社
大阪府門真市大字門真1006番地

(72)発明者　古　山　浩　志
大阪府門真市大字門真1006番地　松下電器産業株式会社内

(72)発明者　井　上　郁　夫
大阪府門真市大字門真1006番地　松下電器産業株式会社内

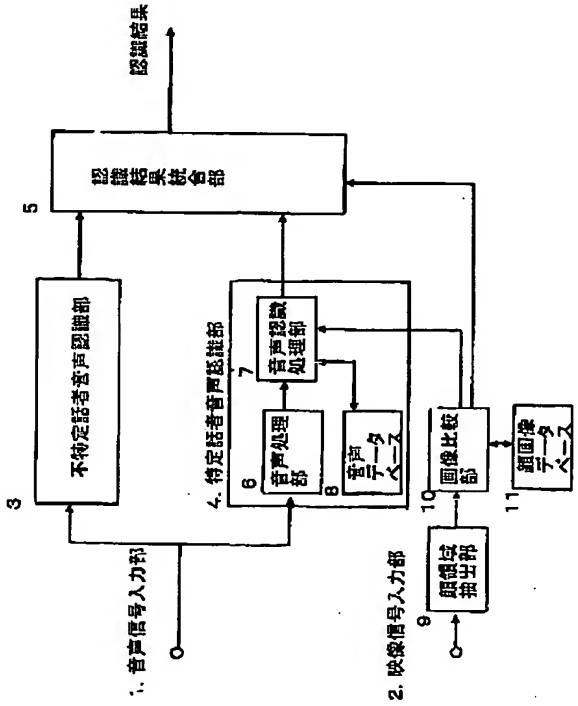
(74)代理人　弁理士　蔵合　正博

(54)【発明の名称】　音声認識装置、話者検出装置及び画像記録装置

(57)【要約】

【課題】　複数の話者に対して、信頼性の高い音声認識を実現する。

【解決手段】　音声信号入力部１と、映像信号入力部２と、複数話者の音声から共通する特徴を抽出して標準パターンを作成し入力音声と標準音声パターンとの類似度を算出する不特定話者音声認識部３と、予め登録された話者の音声と入力音声との類似度を算出する特定話者音声認識部３と、入力映像から顔領域を抽出する顔領域抽出部９と、複数特定話者の顔画像データと話者識別番号とを記録する顔画像データベース１１と、顔領域抽出部９と顔画像データベース１１から入力する画像データとの類似度を出力する画像比較部１０と、不特定話者音声認識部３、特定話者音声認識部４、及び画像比較部１０の出力から統合的類似度を算出し認識結果を出力する認識結果統合部５とで構成される。



【特許請求の範囲】
【請求項１】　話者の特徴的外観の画像を含む映像データを入力する映像入力手段と、　話者の音声データを入力する音声入力手段と、
複数の特定の話者の音声データを、それを特定できる話者識別情報とともに登録し、登録された音声データと入力音声データとの類似度を算出して音声認識を行う特定話者音声認識手段と、
不特定多数の話者の音声データから共通する特徴を抽出して標準パターンを作成、登録して、音声標準パターンと入力音声データとの類似度を算出して音声認識を行う不特定話者音声認識手段と、
複数の話者の特徴的外観の画像を、その話者を特定できる話者識別情報とともに登録し、登録された画像データと前記映像データに含まれる話者の画像データとの類似度を算出する画像認識手段と、
特定話者音声認識手段の出力と不特定話者音声認識手段からの出力と画像認識手段からの出力とを統合して、音声認識結果として単語等を出力する認識結果統合手段とを備えた音声認識装置。
【請求項２】　話者の特徴的外観を含む映像データを入力する映像入力手段と、
話者の音声データを入力する音声入力手段と、
複数の話者の特徴的外観の画像を、その話者を特定できる話者識別情報とともに登録し、登録された画像データと前記映像データに含まれる話者の画像データとの類似度を算出する画像認識手段と、
複数の特定の話者の音声データを、それを特定できる情報とともに登録し、前記画像認識手段で算出された類似度をもとに登録された音声データを絞り込んだ後、その絞り込まれた音声データと入力音声データとの類似度を算出して音声認識を行う特定話者音声認識手段と、不特定多数の話者の音声データから共通する特徴を抽出して標準パターンを作成、登録して、音声標準パターンと入力音声データとの類似度を算出して音声認識を行う不特定話者音声認識手段と、
特定話者音声認識手段の出力と不特定話者音声認識手段からの出力とを統合して、音声認識結果として単語等を出力する認識結果統合手段とを備えた音声認識装置。
【請求項３】　特定話者音声認識手段では、画像認識手段から出力される類似度が閾値を超えた話者の音声データのみに対して、入力音声データとの類似度を算出する請求項１に記載の音声認識装置。
【請求項４】　特定話者音声認識手段では、画像認識手段から出力される類似度のうち最大となる話者の音声データのみに対して、入力音声データとの類似度を算出する請求項１に記載の音声認識装置。
【請求項５】　認識結果統合手段が、画像認識手段から出力される、話者識別情報に対応する話者の画像データと入力映像に含まれる画像データとの類似度をRi、特定話者音声認識手段から出力される、話者識別情報に対応する話者の音声データｊに対する入力音声データとの類似度をR'i,j、不特定話者音声認識手段から出力される、入力音声データと音声データｊとの類似度をR''とするときに、そのRiとR'i,jとR''jとを用いて最適な音声データを出力することを特徴とする請求項１、３、４のいずれかに記載の音声認識装置。
【請求項６】　顔などの唇を含む外観を話者の特徴的外観とし、入力映像から話者の唇の動きを検出する唇動き検出手段を備え、特定話者認識手段では、単位時間あたりの唇の動き量が設定された閾値よりも大きい入力に対してのみ、入力音声データと登録音声データの類似度を算出することを特徴とする請求項１、３～５のいずれかに記載の音声認識装置。
【請求項７】　話者の顔を含む映像データを入力する映像入力部と、話者の音声データを入力する音声入力部と、入力映像から話者の唇の動きを検出する唇動き検出手段と、入力音声から音声レベルを検出する音声レベル検出手段とを備え、
単位時間あたりの唇の動き量と音声レベルが共に設定された閾値を超えている時には、入力映像中に話者の映像が含まれていることを示す話者検出信号を出力する話者検出装置。
【請求項８】　話者の顔を含む映像データを入力する映像入力部と、話者の音声データを入力する音声入力部と、入力映像から話者の唇の動きを検出する唇動き検出手段と、入力音声から音声レベルを検出する音声レベル検出手段とを備え、単位時間あたりの唇の動き量と音声レベルが共に設定された閾値を超えている時には、入力映像中に話者の映像が含まれていることを示す話者検出信号を出力する話者検出装置を具備し、
特定話者認識手段では、話者検出信号が設定された閾値以上となる入力に対してのみ、入力音声データと登録音声データの類似度を算出することを特徴とする請求項１、３～６のいずれかに記載の音声認識装置。
【請求項９】　音声信号入力部と映像信号入力部は、それぞれ映像表示装置の音声信号出力部と映像出力部に接続され、前記映像表示装置の表示対象である出演者を特定できる出演者識別情報を含む、出演者情報を入力する出演者情報入力手段と、出演者情報を記録する出演者情報記録手段と、出演者情報から現在、表示されている出演者を特定し、登録された画像データの中から特定された出演者の画像データを検索する画像検索手段を備え、画像認識部では、検索された出演者の画像と入力した映像信号に含まれる話者の画像との類似度を算出することを特徴とする請求項１、３～６、８のいずれかに記載の音声認識装置。
【請求項１０】　話者の特徴的外観を含む映像データを入力する映像入力部と、複数の話者の特徴的外観の画像を、それを特定できる話者識別情報と共に登録する画像

データベースと、登録された画像データと入力映像に含まれる話者の画像データとの類似度を算出する画像認識手段とを備え、
入力した映像に含まれる画像と登録された画像データの類似度を算出し、登録されているすべての画像データに対する類似度が予め設定された閾値以下の場合は、未登録の話者として新たな話者識別情報とともに画像データベースに記録する画像記録装置。
【請求項１１】　話者の特徴的外観を含む映像データを入力する映像入力部と、複数の話者の特徴的外観の画像を、それを特定できる話者識別情報と共に登録する画像データベースと、登録された画像データと入力映像に含まれる話者の画像データとの類似度を算出する画像認識手段とを備え、入力した映像に含まれる画像と登録された画像データの類似度を算出し、登録されているすべての画像データに対する類似度が予め設定された閾値以下の場合は、未登録の話者として新たな話者識別情報とともに画像データベースに記録する画像記録装置を具備し、未登録の話者を自動的に登録することが可能な請求項１、３～６、８、９のいずれかに記載の音声認識装置。
【請求項１２】　入力した映像に含まれる画像と登録された画像データの類似度を算出し、出力された類似度が予め設定した閾値Ｓ１以上となる登録話者のすべての音声データに対して、特定話者音声認識手段から出力する入力音声データと登録音声データとの間の類似度が予め設定された閾値Ｓ２以下であり、かつ、不特定話者音声認識手段から出力される候補単語等の類似度が予め設定された閾値Ｓ３以上である場合に、該当の話者の未登録音声データとして、それを特定できる話者識別情報とともに入力音声データを記録するための記録手段を有する請求項１１に記載の音声認識装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、映像信号と音声信号を用いて音声認識を行う音声認識装置に関する。
【０００２】
【従来の技術】音声認識方式には、特定話者音声認識方式と不特定話者音声認識方式とがある。特定話者の音声のみを認識する特定話者音声認識方式は、話者の音声を予め登録し、入力音声と登録音声の類似度を算出し、入力音声の認識を行うものである。従って、登録された話者の音声入力に対しては、高い認識率を得ることができるが、話者の音声を登録する作業が必要となる。一方、不特定話者認識方式では、不特定多数の音声から共通する特徴を抽出して標準パターンを作成し、入力音声と音声標準パターンとの類似度を算出し、入力音声の認識を行うものである。従って、話者の音声を登録する煩雑さはないが、特定話者音声認識方式と比較すると認識率は低くなる。
【０００３】このような、特定話者音声認識方式と不特定話者音声認識方式における、それぞれの短所を補うため、特定話者音声認識方式と不特定話者音声認識方式を併用する音声認識装置が考えられている（特開昭63－32596号公報）。
【０００４】この装置は図６に示すように、音声信号を入力する音声信号入力部1、不特定話者音声認識方式により音声認識を行う不特定話者認識部３、特定話者音声認識方式により音声認識を行う特定話者音声認識部４、不特定話者認識部3と特定話者認識部4でそれぞれ求まった認識結果を入力して、類似度の大きな方の認識結果を出力する認識結果統合部5を備えている。また、認識結果が正解と判断されたときには、入力音声の特徴データを特定話者認識用として特定話者認識部４に登録する。
【０００５】このように、従来の音声認識装置では特定話者音声認識方式と不特定話者音声認識方式を併用することにより、音声認識の認識率を高め、また、特定の話者の音声データを自動的に登録することが可能となっている。
【０００６】
【発明が解決しようとする課題】音声認識装置の用途として、例えばパーソナルコンピューター、TVやVTR等、家庭内にある電気製品の機器制御のための入力装置としての利用が考えられるが、家庭内で利用する場合には、ある特定の人物が発する音声に対してのみ高い認識率を有するのでは不十分であり、同居している家族など、複数の人物から発せられる、それぞれの音声入力に対しても高い認識率を維持する必要がある。
【０００７】本発明は、このような要求にこたえるものであり、複数の話者に対しても高い認識率を実現することができる音声認識方式を提供することを目的としている。
【０００８】
【課題を解決するための手段】そこで、本発明の音声認識装置では、話者の特徴的外観（顔など）を含む映像データを入力する映像入力手段と、話者の音声データを入力する音声入力手段と、不特定話者音声認識認識方式により音声認識を行う不特定話者音声手段と、認識を行う話者を含む複数の話者の音声データを蓄積する音声データベースと、特定話者音声認識方式により音声認識を行う特定話者音声認識手段と、入力する映像から話者の顔領域を抽出する顔領域抽出手段と、認識を行う話者を含む複数の話者の顔画像データを蓄積する顔画像データベースと、顔領域抽出手段から出力される顔画像と顔画像データベースに蓄積された顔画像とを比較して、類似度を出力する画像比較手段と、不特定話者音声認識手段と特定話者音声認識手段からそれぞれ出力される認識候補音声と入力音声との間の類似度と画像比較手段から出力される類似度を統合して、最終的な音声認識結果として出力する認識結果統合手段とを備え、不特定話者音声認

識手段から出力される入力音声と認識候補音声の類似度と、画像比較手段から出力される顔画像データベースに登録された話者の顔画像と入力映像に含まれる顔画像の類似度と、特定話者認識手段から出力される音声データベースに登録された話者の音声データと入力音声の類似度を組み合わせて、総合的な類似度から認識結果を出力するようにしている。

【０００９】また、音声データベース、顔画像データベースに複数の話者のデータが登録されている場合には、それぞれの話者に対する顔画像と音声の類似度から、総合的な類似度を算出し認識結果を出力する。

【００１０】従って、登録された話者の顔画像と入力映像に含まれる話者の顔画像の類似度が小さいときには、不特定話者音声認識手段からの出力が認識結果に大きく寄与し、登録された話者の顔画像と入力した話者の顔画像の類似度が大きいときには、その中でも最も類似度の大きな話者に対する、特定話者音声認識手段からの出力が認識結果に大きく寄与するため、複数の特定話者、あるいは不特定の話者から発せられる音声の入力に対して、より信頼性の高い音声認識が可能となる。

【００１１】

【発明の実施の形態】本発明の請求項１に記載の発明は、音声認識装置に、話者の特徴的外観の画像を含む映像データを入力する映像入力手段と、　話者の音声データを入力する音声入力手段と、複数の特定の話者の音声データを、それを特定できる話者識別情報とともに登録し、登録された音声データと入力音声データとの類似度を算出して音声認識を行う特定話者音声認識手段と、不特定多数の話者の音声データから共通する特徴を抽出して標準パターンを作成、登録して、音声標準パターンと入力音声データとの類似度を算出して音声認識を行う不特定話者音声認識手段と、複数の話者の特徴的外観の画像を、その話者を特定できる話者識別情報とともに登録し、登録された画像データと前記映像データに含まれる話者の画像データとの類似度を算出する画像認識手段と、特定話者音声認識手段の出力と不特定話者音声認識手段からの出力と画像認識手段からの出力とを統合して、音声認識結果として単語等を出力する認識結果統合手段とを備えたものであり、登録された話者の顔画像と入力映像に含まれる話者の顔画像の類似度の大小によって出力元の音声認識手段を変えることにより、複数の特定話者、あるいは不特定の話者から発せられる音声の入力に対して、より信頼性の高い音声認識が可能になるという作用を有する。

【００１２】本発明の請求項２に記載の発明は、音声認識装置に、話者の特徴的外観を含む映像データを入力する映像入力手段と、話者の音声データを入力する音声入力手段と、複数の話者の特徴的外観の画像を、その話者を特定できる話者識別情報とともに登録し、登録された画像データと前記映像データに含まれる話者の画像データとの類似度を算出する画像認識手段と、複数の特定の話者の音声データを、それを特定できる情報とともに登録し、前記画像認識手段で算出された類似度をもとに登録された音声データを絞り込んだ後、その絞り込まれた音声データと入力音声データとの類似度を算出して音声認識を行う特定話者音声認識手段と、不特定多数の話者の音声データから共通する特徴を抽出して標準パターンを作成、登録して、音声標準パターンと入力音声データとの類似度を算出して音声認識を行う不特定話者音声認識手段と、特定話者音声認識手段の出力と不特定話者音声認識手段からの出力とを統合して、音声認識結果として単語等を出力する認識結果統合手段とを備えたものであり、顔などを含む映像から話者の顔画像を抽出して、登録された話者の顔画像データベースと照合し、類似度を算出して、特定話者音声認識部、不特定話者音声認識部から出力する音声の類似度との統合的な類似度を算出して認識結果を出力することにより、複数の特定話者の入力に対して、信頼性の高い音声認識を行うことが可能となるという作用を有する。

【００１３】本発明の請求項３に記載の発明は、請求項１記載の音声認識装置において、特定話者音声認識手段では、画像認識手段から出力される類似度が閾値を超えた話者の音声データのみに対して、入力音声データとの類似度を算出するようにしたものである。

【００１４】本発明の請求項４に記載の発明は、請求項１記載の音声認識装置において、特定話者音声認識手段では、画像認識手段から出力される類似度のうち最大となる話者の音声データのみに対して、入力音声データとの類似度を算出するようにしたものである。

【００１５】本発明の請求項５に記載の発明は、請求項１、３、４のいずれかに記載の音声認識装置において、認識結果統合手段が、画像認識手段から出力される、話者識別情報に対応する話者の画像データと入力映像に含まれる画像データとの類似度をＲｉ、特定話者音声認識手段から出力される、話者識別情報に対応する話者の音声データｊに対する入力音声データとの類似度をＲ'ｉ，ｊ、不特定話者音声認識手段から出力される、入力音声データと音声データｊとの類似度をＲ''とするときに、そのＲｉとＲ'ｉ,ｊとＲ''ｊとを用いて最適な音声データを出力するようにしたものである。

【００１６】本発明の請求項６に記載の発明は、請求項１、３～５のいずれかに記載の音声認識装置において、顔などの唇を含む外観を話者の特徴的外観とし、入力映像から話者の唇の動きを検出する唇動き検出手段を備え、特定話者認識手段では、単位時間あたりの唇の動き量が設定された閾値よりも大きい入力に対してのみ、入力音声データと登録音声データの類似度を算出するようにしたものである。

【００１７】本発明の請求項７に記載の発明は、話者検出装置として、話者の顔を含む映像データを入力する映

像入力部と、話者の音声データを入力する音声入力部と、入力映像から話者の唇の動きを検出する唇動き検出手段と、入力音声から音声レベルを検出する音声レベル検出手段とを備えたものであり、単位時間あたりの唇の動き量と音声レベルが共に設定された閾値を超えている時には、入力された映像中に話者の映像が含まれていることを示す話者検出信号を出力するという作用を有する。

【００１８】本発明の請求項８に記載の発明は、請求項１、３～６のいずれかに記載の音声認識装置において、請求項７に記載の話者検出装置と同様の機能を有する話者検出手段を具備し、特定話者認識手段では、話者検出信号が設定された閾値以上となる入力に対してのみ、入力音声データと登録音声データの類似度を算出するようにしたものである。

【００１９】本発明の請求項９に記載の発明は、請求項１、３～６、８のいずれかに記載の音声認識装置において、音声信号入力部と映像信号入力部は、それぞれ映像表示装置の音声信号出力部と映像出力部に接続され、前記映像表示装置の表示対象である出演者を特定できる出演者識別情報を含む、出演者情報を入力する出演者情報入力手段と、出演者情報を記録する出演者情報記録手段と、出演者情報から現在、表示されている出演者を特定し、登録された画像データの中から特定された出演者の画像データを検索する画像検索手段を備えたものであり、画像認識部では、検索された出演者の画像と入力した映像信号に含まれる話者の画像との類似度を算出し、類似度に応じた音声を出力するという作用を有する。

【００２０】本発明の請求項１０に記載の発明は、話者の特徴的外観を含む映像データを入力する映像入力部と、複数の話者の特徴的外観の画像を、それを特定できる話者識別情報と共に登録する画像データベースと、登録された画像データと入力映像に含まれる話者の画像データとの類似度を算出する画像認識手段とを備えたものであり、入力した映像に含まれる画像と登録された画像データの類似度を算出し、登録されているすべての画像データに対する類似度が予め設定された閾値以下の場合は、未登録の話者として新たな話者識別情報とともに画像データベースに記録することによりデータベースの内容をより充実させるという作用を有する。

【００２１】本発明の請求項１１に記載の発明は、請求項１、３～６、８、９のいずれかに記載の音声認識装置において、請求項１０に記載の顔画像装置と同様の機能を有する画像記録手段を具備したものであり、未登録の話者を自動的に登録することによりデータベースの内容をより充実させるという作用を有する。

【００２２】本発明の請求項１２に記載の発明は、請求項１１記載の音声認識装置において、入力した映像に含まれる画像と登録された画像データの類似度を算出し、出力された類似度が予め設定した閾値Ｓ１以上となる登録話者のすべての音声データに対して、特定話者音声認識手段から出力する入力音声データと登録音声データとの間の類似度が予め設定された閾値Ｓ２以下であり、かつ、不特定話者音声認識手段から出力される候補単語等の類似度が予め設定された閾値Ｓ３以上である場合に、該当の話者の未登録音声データとして、それを特定できる話者識別情報とともに入力音声データを記録するための記録手段を備えたものである。

【００２３】以下、本発明の実施の形態について図１から図５を用いて説明する。なお、本発明はこれら実施の形態に何等限定されるものではなく、その要旨を逸脱しない範囲において種々なる態様で実施し得る。

【００２４】（実施の形態１）第一の実施の形態の音声認識装置は、図１に示すように、音声を入力する音声信号入力部１、映像を入力する映像信号入力部２、複数の不特定話者の音声から共通する特徴を抽出して標準パターンを作成し、入力音声と音声標準パターンとの間の類似度を算出して出力する不特定話者音声認識部３、予め登録された話者の音声と入力音声の類似度を算出して出力する特定話者音声認識部４、入力映像から話者の顔領域を抽出する顔領域抽出部９、複数の特定話者の顔画像データを話者の名前、もしくは話者識別コードとともに記録する顔画像データベース１１、顔領域抽出部９と顔画像データベース１１から入力する画像データを比較し類似度を出力する画像比較部１０、不特定話者認識部３、特定話者認識部４、画像比較部１０から出力されるそれぞれの類似度から、類似度がもっとも大きな類似度に対応する単語等を認識結果として出力する認識結果統合部５により構成される。

【００２５】また、特定話者音声認識部４は、入力した音声をスペクトラム分析等により音声の特徴量を抽出する音声処理部６、隠れマルコフモデル等により入力音声と登録音声データとの間の特徴量の類似度を算出して出力する音声認識処理部７、複数の特定話者の音声データを話者の名前、もしくは話者識別コードとともに記録する音声データベース８により構成される。

【００２６】この装置では、音声入力部１に入力した話者の音声は不特定話者音声認識部３と特定話者音声認識部４内の音声処理部6に入力する。音声処理部６の出力は、音声認識処理部７に入力し、音声認識処理部７の出力は、特定話者音声認識部４の出力として出力される。また、音声認識処理部７と音声データベース８は互いに接続されている。

【００２７】映像信号入力部２に入力した話者の顔を含む映像は、顔領域出力部９に入力し、顔領域抽出部９の出力は画像比較部１０に入力し、画像比較部１０の出力は音声認識処理部７に接続される。また、画像比較部１０と画像データベース１１は互いに接続されている。

【００２８】不特定話者音声認識部３、特定話者音声認識部４、画像比較部１０の出力は、認識結果統合部５に

入力し、認識結果統合部５からは認識結果が出力される。なお、映像信号中に含まれる人物の顔領域部分を抽出する手法は公知であり、例えば、第２回画像センシングシンポジウム講演集、Ａ－１、ｐｐ．１～６、「色情報とＧＡを用いた顔画像抽出と個人照合の応用」などに示されている。また、二つの顔画像データを比較する手法も、例えば、電子情報通信学会論文誌、Ｄ－２，Ｖｏｌ．，Ｊ７６－Ｄ－２，Ｎｏ．６，ｐｐ．１１３２～１１３９、「モザイクとニューラルネットを用いた顔画像の認識」などに示されている。また、こうした技術を用いて、顔画像を個人照合に利用することは特願平８－１７０８６６号公報、或いは特願平８－８６１７１号公報などに示されている。この実施の形態の装置においても、これらの技術を用いることが可能であるが、それだけに限定されるものではない。

【００２９】また、音声認識処理部７における類似度の算出には、隠れマルコフモデル以外にニューラルネットワーク等、一般に用いられている他の手法を用いてもよい。顔画像比較部１０では、登録された複数の話者に対して、ｉ番目の話者の顔画像データと顔領域抽出部９から出力する顔画像データとの間の類似度Ｒｉを算出し出力する。特定話者音声認識部４では、登録された複数の話者に対してｉ番目の話者の音声データｊと、入力音声との類似度をＲｉ，ｊを算出して出力する。不特定話者認識部３では、音声データｊに対して複数の不特定話者の音声から共通する特徴を抽出して作成した標準パターンと入力音声との間の類似度Ｒ’’ｊを算出し、出力する。認識結果統合部５では、あらかじめ設定した係数をαとするときに、登録された話者ｉのすべての音声データｊに対して、α・Ri・R'i,jとR''jを算出し、その値が最大となる音声データｊに対応する単語等を認識結果として出力する。

【００３０】ここで、係数αは、音声データベース８に登録した音声データとは別に、音声データベース８に登録された話者を含む複数の話者の音声データを用意し、この音声データの入力に対して認識率が最大となるようあらかじめ設定するものである。このように特定話者音声認識の方式において、登録された話者の顔画像データと入力映像に含まれる話者の顔画像を比較して、話者を特定することにより、信頼性の高い音声認識が可能となる。

【００３１】（実施の形態２）この音声認識装置は図２に示すように、顔領域抽出部９から出力される話者の顔領域映像を入力し、前述の顔領域抽出部９と同じ手法により話者の口唇部分を抽出した映像を出力する唇領域抽出部１２、唇領域抽出部１２から出力される話者の口唇領域映像を入力し、話者の唇の動きを検出する唇動き検出部１３と、実施の形態１と同じく音声処理部６、音声認識処理部７、音声データベース８から構成される特定話者音声認識部７、音声入力信号部１、映像信号入力部２、不特定話者音声認識部３、顔領域抽出部９、画像比較部１０、認識結果統合部５を備えている。

【００３２】この装置では、音声入力部１に入力した話者の音声は不特定話者認識部３と特定話者認識部４内の音声処理部６に入力する。音声処理部６の出力は、音声認識処理部７に入力し、音声認識処理部７の出力は、特定話者音声認識部４の出力として出力される。また、音声認識処理部７と音声データベース８は互いに接続されている。映像信号入力部２に入力した話者の顔を含む映像は、顔領域出力部９に入力し、顔領域抽出部９の出力は画像比較部１０と唇領域抽出部１２に入力し、画像比較部１０の出力は音声認識処理部７に接続される。また、画像比較部１０と画像データベース１１とは互いに接続されている。

【００３３】唇領域抽出部１２の出力は唇動き検出部１３に入力し、唇動き検出部１３の出力は、音声認識処理部７に入力する。不特定話者音声認識部３、特定話者音声認識部４、画像比較部１０の出力は、認識結果統合部５に入力し、認識結果統合部５からは認識結果が出力される。唇動き検出部１３では、唇領域抽出部１２で抽出した口唇領域の映像から、唇上のある着目点の動きベクトルを検出し、単位時間の唇の動きベクトルの平均値が設定した閾値よりも大きいときには

Ｋ＝１

閾値より小さいときには

Ｋ＝０

を出力する。認識結果統合部５では、登録されたすべての話者ｉの音声データｊに対して、

α・Ｋ・Ri・R'i,jとR''j

を算出し、その値が最大となる音声データｊに対応する単語等を認識結果として出力する。

【００３４】このように特定話者音声認識の方式において、入力映像に含まれる話者の唇の動きを検出することにより、入力映像に複数の話者の顔が含まれているときにも、信頼性の高い音声認識が可能となる。

【００３５】（実施の形態３）この装置は図３に示すように、入力した音声のレベルが閾値を超えているかを検出する音声レベル検出部１４と、実施の形態２と同じく、音声信号入力部１、音声処理部６、音声認識処理部７、音声データベース８から構成される特定話者音声認識部７、音声入力信号部１、映像信号入力部２、不特定話者音声認識部３、顔領域抽出部９、画像比較部１０、認識結果統合部５、唇領域抽出部１２、唇動き検出部１３を備えている。

【００３６】この装置では、音声入力部１に入力した話者の音声は不特定話者認識部３と特定話者認識部４内の音声処理部６に入力する。音声処理部６の出力は、音声認識処理部７に入力し、音声認識処理部７の出力は、特定話者音声認識部４の出力として出力される。また、音声認識処理部７と音声データベース８は互いに接続され

ている。映像信号入力部２に入力した話者の顔を含む映像は、顔領域出力部９に入力し、顔領域抽出部９の出力は画像比較部１０と唇領域抽出部１２に入力し、画像比較部１０の出力は音声認識処理部７に接続される。また、画像比較部１０と画像データベース１１は互いに接続されている。唇領域抽出部１２の出力は唇動き検出部１３に入力し、唇動き検出部１３の出力は、音声認識処理部７に入力する。また、音声入力部１は音声レベル検出部１４にも接続し、音声レベル検出部１４の出力は音声認識処理部７と接続されている。

【００３７】不特定話者音声認識部３、特定話者音声認識部４、画像比較部１０の出力は、認識結果統合部５に入力し、認識結果統合部５からは認識結果が出力される。また、唇動き検出部１３では、唇領域抽出部１２で抽出した口唇領域の映像から、唇の動きを検出し、単位時間の唇の動きの平均値が設定した閾値よりも大きいときには

Ｋ＝１

閾値よりも小さいときには

Ｋ＝０

を出力する。音声レベル検出部１４では、単位時間の音声レベルの平均値が設定した閾値よりも大きいときには

Ｌ＝１

閾値よりも小さいときには

Ｌ＝０

を出力する。認識結果統合部５では、登録されたすべての話者ｉの音声データｊに対して、

$\alpha \cdot K \cdot L \cdot Ri \cdot R'i,j$ と $R''j$

を算出し、その値が最大となる音声データｊに対応する単語等を認識結果として出力する。

【００３８】このように特定話者音声認識の方式において、入力映像に含まれる話者の唇の動きを検出することにより、入力映像に複数の話者の顔が含まれているときにも、より信頼性の高い音声認識が可能となる。

【００３９】（実施の形態４）この装置は図４に示すように、ＴＶ番組の出演者の名前のデータを含む番組表を入力する番組表入力部１５、番組表を記録する番組表記録部１６、番組表データと現在の時刻を比較し、現在、放送されているＴＶ番組の出演者を特定して出演者の名前を出力する出演者名検出部部１７、出演者名検出部１７から出力する出演者の名前から顔画像データベース１１を検索して、検索した顔画像を出力させる画像検索部１８と、実施の形態１と同じく音声処理部６と音声データベース８と音声認識処理部７から構成される特定話者認識部４と、音声信号入力部１、映像信号入力部２と、不特定話者音声認識部３と、顔領域抽出部９と、画像比較部１０と、認識結果統合部５を備えている。

【００４０】この装置では、番組表入力部１５に入力した番組表データは、番組表記録部１６に入力し、記録される。出演社名検出部１７は、番組表記録部１６と画像検索部１８に接続され、画像検索部１９と顔画像データベース１１とは互いに接続されている。また、音声入力部１に入力した話者の音声は不特定話者認識部３と特定話者認識部４内の音声処理部６に入力する。音声処理部６の出力は、音声認識処理部７に入力し、音声認識処理部７の出力は、特定話者音声認識部４の出力として出力される。また、音声認識処理部７と音声データベース８は互いに接続されている。映像信号入力部２に入力した話者の顔を含む映像は、顔領域出力部９に入力し、顔領域抽出部９の出力は画像比較部１０に入力し、画像比較部１０の出力は音声認識処理部７に接続される。また、画像比較部１０と画像データベース１１とは互いに接続されている。

【００４１】不特定話者音声認識部３、特定話者音声認識部４、画像比較部１０の出力は、認識結果統合部５に入力し、認識結果統合部５からは認識結果が出力される。また、音声信号入力部１と映像信号入力部２は、それぞれＴＶ受信機の音声信号出力端子と映像出力端子に接続する。出演者名検出部１７は、番組表記録部１６に記録されているＴＶ番組表データと現在の時刻から、現在放送されているＴＶ番組の出演者を特定し、特定した出演者の名前データを出力する。画像検索部１８は、時刻比較部１７から出力する出演者の名前データをもとに画像検索データベース１１から、出演者の顔画像を検索して、画像比較部１２へと出力させる。顔画像比較部１０では、ＴＶ番組の出演者と特定された複数の話者に対して、ｋ番目の話者の顔画像データと入力映像に含まれ顔領域抽出部９から出力する顔画像データとの間の類似度ＲＫを算出し出力する。特定話者音声認識部４では、出演者と特定された複数の話者に対してｋ番目の話者の音声データｊと、入力音声との類似度を$R'k,j$を算出して出力する。不特定話者認識部３では、音声データｊに対して複数の不特定話者の音声から共通する特徴を抽出して作成した標準パターンと入力音声との間の類似度$R''j$を算出し、出力する。音声認識認識統合部では、あらかじめ設定した係数をαとするときに、出演者として特定されたすべての話者ｋの音声データｊに対して、

$\alpha \cdot Rk \cdot R'k,j$ と $R''j$

を算出し、その値が最大となる音声データｊに対応する単語等を認識結果として出力する。

【００４２】このように特定話者音声認識の方式において、番組表データを本にＴＶ番組の出演者を特定し、特定された出演者の顔画像データとＴＶ番組映像に含まれる話者の顔画像を比較して、話者を特定することにより、より信頼性の高い音声認識が可能となる。なお、本実施の形態では映像表示装置としてＴＶ受信機を例にあげて説明したが、ＶＴＲ・ビデオ等の映像を表示するものであれば何でも構わない。

【００４３】（実施の形態５）この装置は図５に示すように、複数の特定話者の音声データを話者の名前、もし

くは話者識別コードとともに記録し、かつ新規に音声データを追加記録する機能を有するする音声データベース８、複数の特定話者の顔画像データを話者の名前、もしくは話者識別コードとともに記録し、かつ新規に顔画像データを追加記録する機能を有する顔画像データベース１１、音声データベース８、顔画像データベース１１にデータを追加記録するための制御を行う記録制御部１９を備え、第一の実施の形態と同じく、音声処理部６、音声データベース８、音声認識処理部７により構成される特定話者音声認識部４と、音声信号入力部１と、映像信号入力部２と、不特定話者音声認識部３と、認識結果統合部５と、顔領域抽出部９と、顔画像データベース１１により構成される。

【００４４】この装置では、音声入力部１に入力した話者の音声は不特定話者認識部３と特定話者認識部４内の音声処理部６に入力する。音声処理部６の出力は、音声認識処理部７に入力し、音声認識処理部７の出力は、特定話者音声認識部４の出力として出力される。また、音声認識処理部７と音声データベース８は互いに接続されている。映像信号入力部２に入力した話者の顔を含む映像は、顔領域出力部９に入力し、顔領域抽出部９の出力は画像比較部１０に入力し、画像比較部１０の出力は音声認識処理部７に接続される。また、画像比較部１０と画像データベース１１とは互いに接続されている。不特定話者音声認識部３、特定話者音声認識部４、画像比較部１０の出力は、認識結果統合部５に入力し、認識結果統合部５からは認識結果が出力される。記録制御部２には、画像比較部１０、不特定話者音声認識部３、特定話者音声認識部４の出力が接続され、記録制御部２０の出力は音声データベース８と顔画像データベース１１に接続される。

【００４５】顔画像比較部１０では、登録された複数の話者に対して、ｉ番目の話者の顔画像データと入力映像に含まれ顔領域抽出部９から出力する顔画像データとの間の類似度Ｒｉを算出し出力する。顔画像データベース１１に登録されたすべての話者ｉに対して、画像比較部１０の出力Ｒｉが、あらかじめ設定された閾値以下の場合は、記録制御部１９は、新しい識別コードを付与して、顔領域抽出部９の出力を顔画像データベース１１に記録するよう制御する。

【００４６】特定話者音声認識部４では、登録された複数の話者のうち画像比較部の出力Ｒｉがあらかじめ設定された閾値Ｓ１以上となる話者に対してｉ番目の話者の音声データｊと、入力音声との類似度をＲ'ｉ,ｊを算出して出力する。不特定話者認識部３では、音声データｊに対して複数の不特定話者の音声から共通する特徴を抽出して作成した標準パターンと入力音声との間の類似度Ｒ''ｊを算出し、出力する。音声データｊに対して、特定話者音声認識部４の出力Ｒ'ｉ,ｊが、あらかじめ設定した閾値Ｓ２よりも小さく、かつ、不特定話者音声認識部３の出力Ｒ''ｊがあらかじめ設定した閾値Ｓ３よりも大きい場合には、記録制御部１９は、音声データ番号ｊ、話者の名前、あるいは話者識別コードとともに入力音声を音声データベースに記録するよう制御する。

【００４７】このように入力音声の話者が未登録の場合は話者の顔画像データと音声データを、話者は登録済みで音声データが未登録の場合は音声データを自動的に追加記録することが可能となる。なお、上記実施の形態１から５では、認識結果統合部５の入力として画像比較部１０からの出力を要件としている（方式１）が、それを必須とはしない不特定話者音声認識部３及び音声認識処理部７の２つの出力を入力（方式２）としても一向に構わない。

【００４８】方式２による方法では、主に以下の場合に有効である。すなわち、（１）　一般に画像処理は負荷が大きいためそれを軽減する対策として、音声認識処理部７の処理対象を画像比較部１０で類似度の大きいものに絞ることにより、音声認識処理部７の処理負荷を少しでも軽減したい場合、（２）　音声データベース８のデータが膨大な為、（画像処理の負荷の大小に関係なく）音声認識処理部の負荷を軽減したい場合、などである。

【００４９】一方、方式１では上記実施の形態１から５の内容に加え、方式２と比較した場合、例えば以下のような場合に有効である。すなわち、（３）　例えば登録された顔画像が正面である場合には、画像比較の対象として横顔等が入力されると、正面から捉えた顔（外観的特徴）でない為、その類似度（画像比較部１０からの出力）の信頼性はやや低下する。そのような場合、方式２のように音声データベース８を前記類似度で絞り込むと音声認識処理部７の出力の信頼性を低下させる可能性があるので、認識結果統合部５での統合化処理の優先度として、不特定話者音声認識部３及び音声認識処理部７の出力（絞り込みを行わない出力）を優先しつつ画像比較部１０の出力も有効活用したい場合、などである。

【００５０】もちろん、上記方式１、方式２を（自動）切り替えするようにすれば、様々な利用形態に対応したより信頼性の高い認識結果（認識結果統合部５の出力）が得られることは言うまでもない。

【００５１】

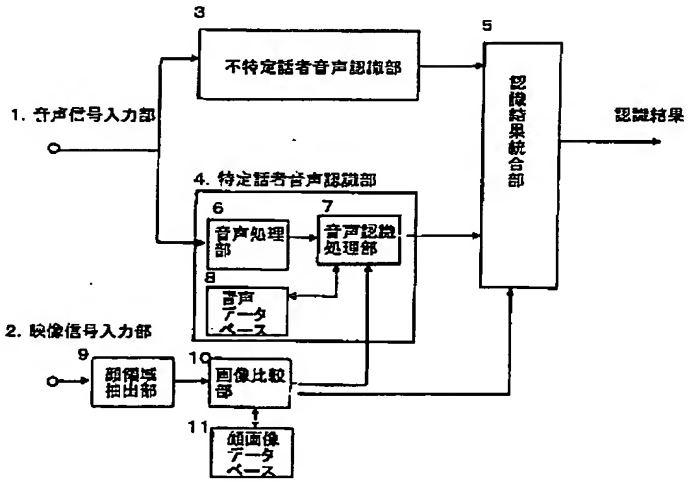
【発明の効果】以上の説明から明らかなように、本発明の音声認識装置は話者の特徴的外観として、例えば顔などを含む映像から話者の顔画像を抽出して、登録された話者の顔画像データベースと照合し、類似度を算出して、特定話者音声認識部、不特定話者音声認識部から出力する音声の類似度との統合的な類似度を算出して認識結果を出力することにより、複数の特定話者の入力に対して、信頼性の高い音声認識を行うことが可能となる。

【図面の簡単な説明】

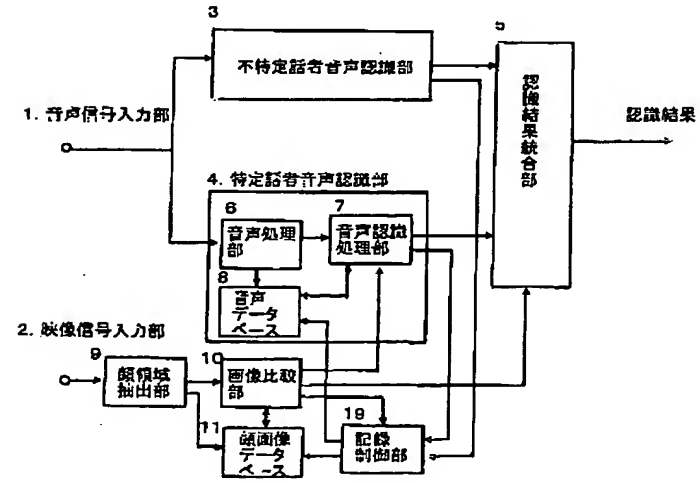
【図１】本発明の第１の実施形態における音声認識装置の概略構成を示すブロック図

【図２】本発明の第２の実施形態における音声認識装置の概略構成を示すブロック図
【図３】本発明の第３の実施形態における音声認識装置の概略構成を示すブロック図
【図４】本発明の第４の実施形態における音声認識装置の概略構成を示すブロック図
【図５】本発明の第５の実施形態における音声認識装置の概略構成を示すブロック図
【図６】従来の音声認識装置の概略構成図
【符号の説明】
１　音声信号入力部
２　映像信号入力部
３　不特定話者音声認識部
４　特定話者音声認識部
５　認識結果統合部
６　音声処理部
７　音声認識処理部
８　音声データベース
９　顔領域抽出部
１０　画像比較部
１１　画像データベース
１２　唇領域抽出部
１３　唇動き検出部
１４　音声レベル検出部
１５　番組表入力部
１６　番組表記録部
１７　出演者名検出部
１８　画像検索部
１９　記録制御部

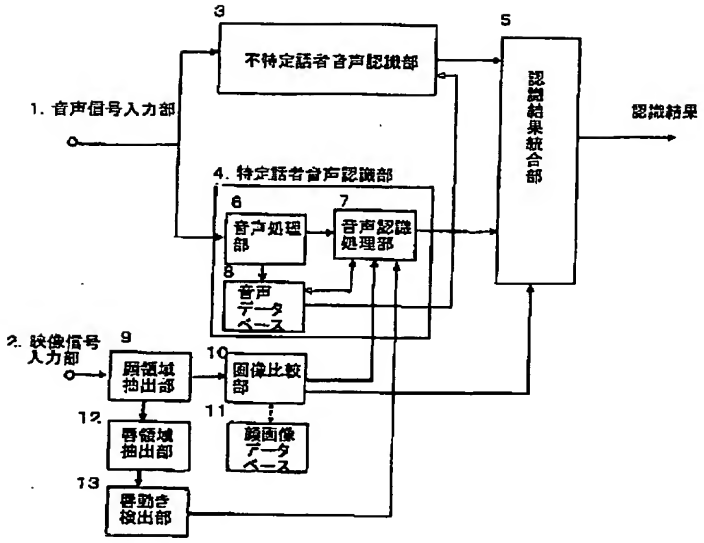
【図１】



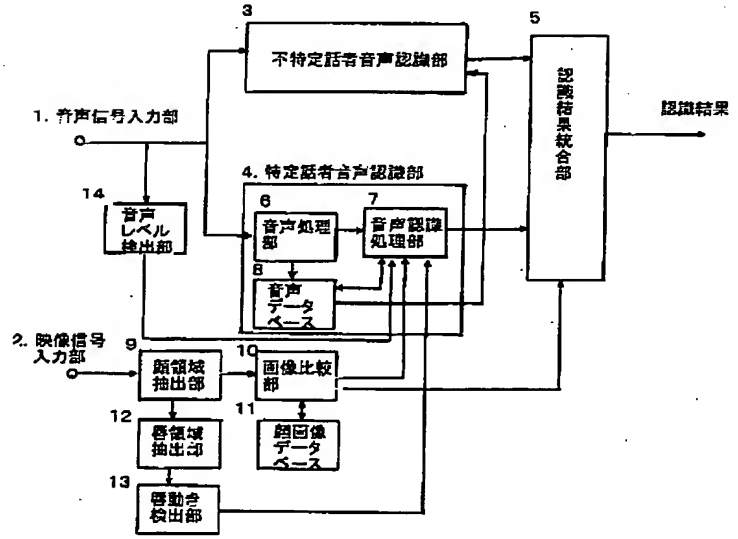
【図５】



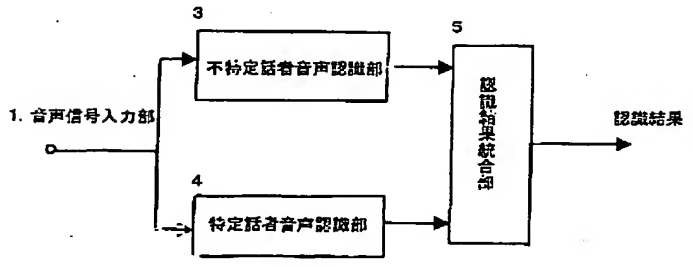
【図２】


3
不特定話者音声認識部
5
認識結果統合部
認識結果
1. 音声信号入力部
4. 特定話者音声認識部
6
音声処理部
7
音声認識処理部
8
音声データベース
2. 映像信号入力部
9
顔領域抽出部
10
画像比較部
11
顔画像データベース
12
唇領域抽出部
13
唇動き検出部

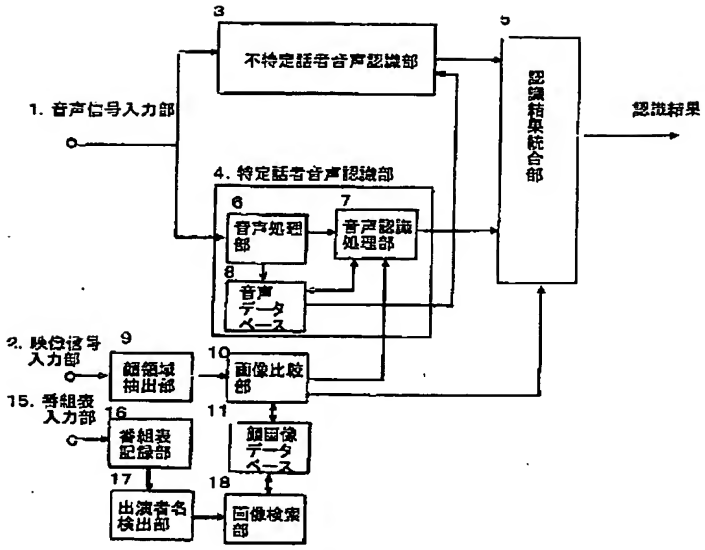
【図３】


3
不特定話者音声認識部
5
認識結果統合部
認識結果
1. 音声信号入力部
4. 特定話者音声認識部
14
音声レベル検出部
6
音声処理部
7
音声認識処理部
8
音声データベース
2. 映像信号入力部
9
顔領域抽出部
10
画像比較部
11
顔画像データベース
12
唇領域抽出部
13
唇動き検出部

【図６】


3
不特定話者音声認識部
5
認識結果統合部
認識結果
1. 音声信号入力部
4
特定話者音声認識部

【図４】



フロントページの続き

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ１０Ｌ　5/06		Ｇ０６Ｆ　15/62	３８０